When does Partial Priority Improve Revenue?

Zhouzi Li[1*], Mor Harchol-Balter[1] and Alan Scheller-Wolf[1]

[1*]Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, 15213, PA, USA.

*Corresponding author(s). E-mail(s): zhouzil@andrew.cmu.edu;
Contributing authors: harchol@cs.cmu.edu; awolf@andrew.cmu.edu;

Abstract

Priority queues have long been used to increase revenue by exploiting the fact that time-sensitive customers are willing to pay for shorter waiting times. This fact begs the question: Can one make even more revenue by relaxing the strictness of the priority policy? This paper answers this question under the unobservable queue setting, where customers are heterogeneous in their time-sensitivity; specifically the time-sensitivity of customers is allowed to follow an arbitrary distribution.

In this paper we prove necessary and sufficient conditions under which partial priority can increase the revenue. Specifically, we find a surprising result: Although partial priority offers much more flexibility than strict priority, partial priority only increases revenue if there are two additional constraints on the service provider, one setting a maximum price and the other setting a maximum waiting time. In the absence of either of these constraints, we prove that strict priority maximizes revenue. Finally, in situations where partial priority increases the revenue, we analytically characterize the amount of improvement.

Keywords: Hybrid, revenue maximization, priority queue, achievability region, bounded wait times, bounded price, time-sensitivity

1 Introduction

The concept of generating revenue by selling queue priority is well-established, particularly when serving customers with different degrees of time sensitivity (cost for

waiting). For example, time-sensitive customers may pay for a Pre-Check to join a priority line at the security check in airports, for expedited passport service, or expedited manufacturing of a critical good.

In practice, priority is typically implemented as a *strict priority* system, where first-class customers always receive service before all second-class customers.[1] Numerous studies have examined revenue-maximizing mechanisms and pricing strategies under strict priority (e.g. [1–6]). These studies have shown that one can leverage the different time sensitivity of customers to increase revenue.

This begs the question of whether one can make even more money by relaxing the strictness of the priority. For example, imagine that class 1 customers get priority with some probability q, say 70%, and class 2 customers get priority with probability $1-q$. We refer to this policy as Hybrid(q) (see Section 3.1 for more details). Hybrid(q) falls within the general class of *partial priority* policies which offer more flexibility for the service provider.

To understand the added flexibility attainable from partial priority, we look at Figure 1. Let $\mathbf{E}[W_1]$, respectively $\mathbf{E}[W_2]$, denote the expected waiting time of first-class and second-class customers. Then, under strict priority, the set of all possible expected waiting time pairs under all possible arrival rates is the blue shaded region in Figure 1. In contrast, in yellow we see the many additional pairs that are possible under partial priority. Following [7], we refer to all pairs (blue + yellow) as the *Achievability Region* of a partial priority system.[2]

While many papers do not explicitly restrict themselves to strict priority (e.g., [9–20]), they also do not specify *whether* non-strict priority is actually helpful. Given that implementing partial priority can be more challenging than implementing strict priority, it is important to look at the following questions:

[1] Throughout this paper, we discuss non-preemptive policies, which means when customers are receiving service, they cannot be interrupted by other customers.

[2] Previous works (e.g., [7, 8]) only talk about the achievability region for fixed arrival rates. We generalize the definition to all possible arrival rates, and the proof is in Appendix B.

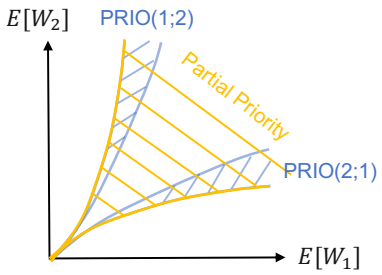


Fig. 1: Achievability region of strict priority and partial priority.

Does the extra flexibility afforded by partial priority policies bring in more revenue? If so, what is the amount of the increase in revenue?

Of all papers mentioned above, only two touch on the first question and none of them answers the second. Moreover, their answers are limited to narrow settings. Specifically, Hassin et al. [16] only demonstrate the benefit of partial priority for revenue maximization numerically, in the special case where the service provider is limited to an exogenous fixed price for each of the two classes. And in [10], customers in both classes have the *same* time-sensitivity, although they are different in how many times they want to use the system. See Section 2 for more details on prior works.

The goal of this paper is to answer both questions under a more general model, in which customers' time sensitivities are drawn from an arbitrary distribution.

1.1 Our model

We model the common situation where there is a social amenity that attracts a steady stream of people. Everyone joins the queue, but people can choose to pay extra for priority within the queue.

A typical example is certain popular attractions at Disney World: after entering Disney World, customers get access to free "standby" lines for attractions. They also have the option to purchase a priority pass called a *lightning pass* for an added cost. At each attraction, customers possessing a lightning pass for the attraction enter a

fast queue which has *strict priority* over the standby queue ([21]). When making their decision to buy a lightning pass, Disney customers do not get to see the queue at each attraction [22]. Thus customers are in an *unobservable* setting, where they only have historical estimates of mean waiting times with and without priority.

Consistent with the above Disney World example, we assume that customers arrive into the system (i.e., an attraction at Disney World) according to a Poisson process with rate λ. The customers only differ in their time-sensitivity, modeled by their *impatience factor*, C, which can follow any given distribution. The service provider needs to serve all the customers, but it wants to leverage the fact that some customers are more impatient to generate revenue. Thus, the service provider selects a priority policy (not necessarily strict priority) and sells the priority (access to queue 1) for price \$. Without loss of generality, we assume the price to enter queue 2 is free: Having a non-zero general entrance fee, as in the Disney example, adds a constant to the total revenue and does not change the optimization problem.

Again consistent with the Disney World example, our model assumes that the state of the queues is *unobservable* to the customers, meaning that the customers can only make decisions based on the expected waiting time pair, $\mathbf{E}[W_1], \mathbf{E}[W_2]$. More specifically, each customer chooses to buy priority or not based on the price, \$, the expected waiting time at each queue, and her own impatience factor: A customer with impatience factor c is willing to buy priority iff

$$c \cdot (\mathbf{E}[W_1] - \mathbf{E}[W_2]) > \$.$$

Let λ_1 denote the arrival rate of customers choosing to buy the priority (and enter queue 1), and $\lambda_2 = \lambda - \lambda_1$ denote the arrival rate of queue 2. The goal of the service

provider is to maximize their revenue rate where:

$$\text{Revenue rate} = \lambda_1 \cdot \$.$$

The model is illustrated in Figure 2, and we provide more details in Section 3.

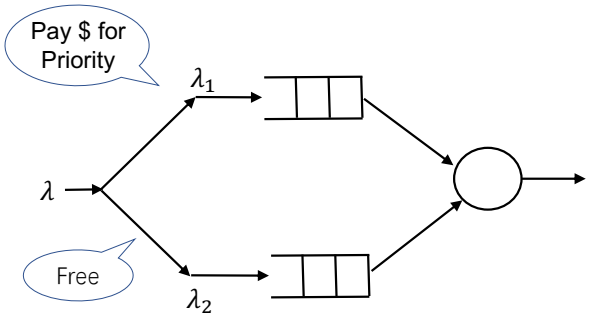


Fig. 2: Illustration for the model.

1.2 Common restrictions within our setting

Absent any restrictions on our setting, it is straightforward to show that strict priority is superior to partial priority. To see why, observe that queue 2 customers do not contribute to the revenue. Thus any (partial) priority given to queue 2 customers is wasted. Therefore, a strict priority policy is optimal.

In practice, however, there are a few common restrictions within our setting. First of all, the service provider typically needs to ensure that the expected waiting time for class two customers, $\mathbf{E}[W_2]$, does not get unbearably long. For example, Disney seems to try to limit $\mathbf{E}[W_2] < \overline{W}$ by limiting the number of people who can buy lightning passes within a given time-period for a given attraction (hence controlling λ_1). It is unknown whether strict priority is still optimal in maximizing revenue given this restriction.

Secondly, there is typically an upper limit, $\overline{\$}$, on how much the service provider charges. For example, in an effort to maintain its family-friendly image, Disney keeps the price of its lighting pass under $\overline{\$} = 30$ dollars per attraction, despite many customers likely being willing to pay more (an example of the media pressure on Disney over costs, including for lightning passes, is given in [23]). Again, it is unknown whether strict priority is still optimal given this restriction.

The literature on pricing for queueing includes many different models. As such, our restrictions on the maximum expected waiting time, $\overline{W}$, and the price cap, $\overline{\$}$, may show up in the literature in different forms. For example, in papers that assume customers have a "utility" for service, the fact that customers' net utility must be positive functions as a restriction, similarly to our $\overline{W}$ or $\overline{\$}$ restrictions. In our model, since all customers enter the system, there is no need for a utility and thus we can be more explicit about our $\overline{W}$ and $\overline{\$}$ restrictions.

1.3 Our results and contributions

There are many reasons to believe that partial priority should increase revenue. For example, by offering class 2 customers slightly more priority, we can lessen their waiting time, making it easier to adhere to the $\overline{W}$ restriction, thus allowing us to admit more class 1 paying customers than under strict priority.

Surprisingly, we prove that if *either* of the above two restrictions ($\overline{W}$ or $\overline{\$}$) are absent, then strict priority will maximize revenue (see Corollary 1 and Remark 1). That is, *both* the restriction $\overline{W}$ *and* the $\overline{\$}$ are needed for partial priority to help. Essentially, we prove a necessary and sufficient condition on the system parameters under which revenue is improved via partial priority (Theorem 1). This effectively tells us how tight the restrictions must be for partial priority policies to increase revenue. Moreover, when partial priority does help increase revenue, we provide Theorem 2

that characterizes the ratio and the absolute amount of revenue improvement. These two results are the main contributions of our paper.

2 Prior Work

Our work generally fits within a research area called "pricing for queueing." This is an area where time-sensitive customers are willing to pay for priority. We start in Section 2.1 by reviewing all the related work on pricing for queueing.

Our work also fits within an area called achievability region analysis. This is an area where one tries to understand what waiting times are possible/impossible. We discuss prior work on achievability region analysis in Section 2.2.

2.1 Pricing for Queueing

Research on charging people prices in exchange for shorter queueing times is not new. The field was perhaps started by Naor in 1969, [24]. Hassin and Haviv's highly cited book followed in 2003 ([4]); it provides an excellent survey of this field.

The general setting is one in which customers arrive and are charged different amounts to enter different queues, where some queues have a higher priority of being served than others. In some works, it is assumed that the arriving customers can observe the queue lengths when deciding which queue to join [6, 25–28] In contrast, our paper focuses on the *unobservable setting* where arriving customers cannot see the queues and need to make their decisions based only on long-run expected waiting times (and prices). For the rest of this section, we limit our attention to the unobservable setting.

Within the **unobservable setting**, several streams of work exist.

Strict Priority: the largest stream of work studies pricing under a strict priority queueing policy, where customers in one queue always have 100% priority over those in another queue. Papers ranging over several decades have worked on deriving the

optimal pricing mechanism in a strict priority system, e.g., [1–3, 5, 29] While well-studied, this stream of work is less related to our paper, and we lay more emphasis on related work under partial priority.

Partial Priority with DPS queueing policy: Discriminatory Processor Sharing (DPS) is one of the most typical partial priority policies. Under the DPS policy, the server is time-shared between two queues in a *preemptive* manner, where each queue gets some fraction of the server. The goal of most of the papers on pricing with DPS is *not* to maximize revenue, but rather to study the equilibrium behavior of customers [17, 18]). An exception is [16], in which Hassin and Haviv characterize the optimal DPS policy to maximize the revenue when given an exogenous fixed price for each of the two queues. They provide *numerical* evidence that DPS can bring more revenue than strict priority policy without analytical proof.

It is interesting that the authors of [16] first prove theorems with respect to equilibrium behaviors in the setting where prices are not fixed, but change to fixed prices setting when talking about maximizing the revenue. Our paper provides an intuitive explanation for this fact: Unless the prices are constrained, partial priority cannot beat strict priority in maximizing the revenue. Observe that fixing the prices is more restrictive than having a price cap.

Partial Priority with Accumulated priority policy: The accumulated priority policy was first raised and analyzed in [30]. Under the accumulated priority policy, every customer enters the system with priority 0. While waiting, each customer accumulates her priority with rate b, and customers in the high-priority classes accumulate priority with higher rate. At each moment of time, the policy serves the customer with the highest priority. Accumulated priority is a type of partial priority, in that low priority class customers (the ones who accumulate priority with a lower rate) may get ahead of high priority class customers if they have waited for a longer time.

Only recently, accumulated priority has been considered in the area of pricing for queueing [19, 20]. However, these papers assume that the price must have a linear relation with the accumulating rate. For example, in [19], the price for having the accumulating rate b is also b. In this way, while interesting results are derived, this stream of papers is less related to our paper where the price does not need to follow a given function of the priority.

Partial Priority with unspecified queueing policy: There is also a large stream of work which does not specify the queueing policy. Instead, their goal is to specify what the expected waiting times would be in an optimal "price-delay menu" that could be offered to customers. They do this by leveraging a body of research on achievability regions (see Section 2.2 below); examples include [9–15], and Afeche established a standard framework in [15]. However, most of these papers do not compare with strict priority.

An exception is [10], in which the authors analytically characterize cases where the optimal policy is strict priority, First Come First Serve (FCFS), or some partial priority. However, in [10] customers have the same time sensitivity and only differ in their "usage rate"; thus FCFS might be the optimal policy. In contast, in our setting where customers can have different time sensitivity, FCFS can never be optimal, and the comparison is more on the priority side. Moreover, another difference is that our paper characterizes *how much* partial priority can increase the revenue, while the characterization in [10] is only qualitative.

Partial Priority in other settings: The idea of a "partial priority" has come up in other settings that do not involve pricing or queueing. Some examples include partial priority in networks [31–34], and inventory rationing with partial priority [35, 36]. While these papers show that partial priority can be useful in many settings, they neither consider pricing nor strategic customers.

2.2 Achievability Region

Our paper also contributes to a relatively understudied but important research area on achievability regions. This concept was first introduced by Coffman in 1980, [7], with subsequent analysis by Federgruen and Groenevelt [8]. Importantly, all works in this literature assumed a default partial priority policy proposed in [7]: the service provider randomly selects a strict priority to implement at the beginning of each busy period.

Our work extends this area of research in two ways: (i) We analyze the achievability region of a new policy, Hybrid; and (ii) unlike previous works, which primarily focus on fixed arrival rates for each job type, we explore the entire space of possible arrival rates. As a result, none of the existing papers have captured the blue "tornado"-shaped achievability regions of strict priority that we show in Figure 1, nor the broad yellow region spanned by partial priority policies.

3 Model

In this section, we describe our model in Section 1 in detail. We will first define the "Hybrid" policy which is a canonical representative of partial priority policies, then describe our model in detail. Note that although the model is stated for the Hybrid policy, all theorems hold generally for partial priority policies.

3.1 Hybrid policy

The Hybrid policy is defined through a parameter $q \in [0, 1]$: Whenever the server is free and there exist customers in the queue, $Hybrid(q)$ flips a coin which comes up heads with probability q and tails otherwise.

- If the coin comes up heads, then the server takes a customer from queue 1. If there are no customers in queue 1, then the server takes a customer from queue 2, assuming one exists.

- If the coin comes up tails, then the server takes a customer from queue 2. If there are no customers in queue 2, then the server takes a customer from queue 1, assuming one exists.

Note that if the system is empty, Hybrid(q) will wait for the next arrival, serve that arrival, and flip its coin at that time.

Observe that Hybrid differs from the policy in [7] in that a decision is made after each customer service, rather than only at the end of a busy period. In this way, Hybrid offers a more fine-grained partial priority, while still enjoying the same achievability region as the other partial priority policies.

We defer the analysis of the achievability region of Hybrid (i.e., the proof of Figure 1) to Appendix B.

3.2 Our Model

In this section we describe our model in detail.

System: There is a single server which serves customers, who arrive according to a Poisson process with rate λ. All customers must be served (no abandonment). Customers are divided into two queues (queue 1 and queue 2). Customers have to pay a price, \$, to enter queue 1. Entering queue 2 is free. When the server is free, the server serves a customer according to the Hybrid(q) policy. If $q = 1$, this corresponds to strict priority.

Customers: All customers have i.i.d. service requirement (service time need) drawn from the distribution denoted by random variable S, where the mean of S is $\mathbf{E}[S] = \frac{1}{\mu}$ and $\mathbf{E}[S^3]$ exists (while the assumption on $\mathbf{E}[S^3]$ is only needed for Proposition 6, that proposition is key to the rest of the paper). Let $\rho := \frac{\lambda}{\mu} < 1$ denote the total load of the system.

Customers are time-sensitive, meaning that they are willing to pay for shorter waiting time. Specifically, there is an impatience factor, C, associated with each customer, where C is a random variable specified in dollars per unit waiting time. Thus a customer with impatience C, who experiences waiting time W, will experience a cost of $C \cdot W$ dollars. We make a mild assumption that the tail of C, denoted by $\bar{F}_C(\cdot)$, is continuous and invertible.

We say that a customer is *class 1* if she decides to buy priority (i.e., join queue 1). Those customers who choose not to buy priority are *class 2*. Let λ_1 denote the arrival rate of class 1 customers and let λ_2 denote the arrival rate of class 2 customers.

Waiting Times: The waiting time of a customer is the time from when the customer arrives to the system until the customer first receives service. We use the random variable W_1 to denote the *waiting time* of class 1 customers. Likewise W_2 will denote the waiting time of class 2 customers.

Customers are willing to pay for priority if and if only if the expected value of the reduction in their waiting time from buying priority is at least \$, the price of joining the priority queue. Mathematically, a customer with impatience factor $C = c$ is willing to buy priority iff

$$c(\mathbf{E}[W_2] - \mathbf{E}[W_1]) \geqslant \$. \tag{1}$$

We assume that there is a restriction on the maximum mean waiting time of class 2 customers; this upper limit is $\overline{W}$. Thus we are restricted to:

$$\mathbf{E}[W_1] < \mathbf{E}[W_2] \leqslant \overline{W}. \tag{2}$$

Price Cap: The service provider has to set the price no more than the price cap $\overline{\$}$, i.e.,

$$\$ \leqslant \overline{\$}. \tag{3}$$

Revenue: The *revenue* that the service provider brings in per unit time is defined as

$$\text{Revenue} := \lambda_1 \cdot \$.$$

Decision Variables: The service provider can control $\$, \lambda_1$ and the parameter q for the queueing policy Hybrid(q) to maximize its revenue. The service provider however is required to adhere to waiting times and prices that are not excessive (the particular values of $\overline{W}$ and $\overline{\$}$ are externally provided).

Customers' Incentive inequality: Observe that the fraction of customers who buy priority, namely $\frac{\lambda_1}{\lambda}$, is upper-bounded by the fraction who *want* to buy priority (i.e., (1) holds). Mathematically this says:

$$\frac{\lambda_1}{\lambda} \leqslant \bar{F}_C\left(\frac{\$}{\mathbf{E}[W_2] - \mathbf{E}[W_1]}\right). \tag{4}$$

In our model, the service provider can limit λ_1 by controlling the number of priority tickets sold. There is a slightly different model where the service provider cannot limit λ_1, i.e., whenever customers want to buy the priority, there is no way of stopping them. In this case, the inequality (4) is an equality. We discuss this variation of our model in Section 4.3.

Optimization Problem: The service provider's optimization problem can be formulated as follows:

$$\boxed{\begin{array}{ll} \underset{\lambda_1, q, \$}{\text{maximize}} & \$ \cdot \lambda_1 \\ \text{s.t.} & \dfrac{\lambda_1}{\lambda} \leqslant \bar{F}_C\left(\dfrac{\$}{\mathbf{E}[W_2] - \mathbf{E}[W_1]}\right), \\ & \mathbf{E}[W_2] \leqslant \overline{W}, \\ & \$ \leqslant \bar{\$}, \\ & 0 \leqslant q \leqslant 1. \end{array}} \tag{5}$$

Traffic Assumptions: Let w_{FCFS} denote the mean waiting time if all customers are served in First-Come-First-Served order (FCFS). Mathematically, $w_{FCFS} := \frac{\rho \mathbf{E}[S_e]}{1-\rho}$. To eliminate uninteresting cases, we make the following assumptions on $\overline{W}$:

1. $\overline{W} > w_{FCFS}$. If $\overline{W} < w_{FCFS}$, no scheduling policy can meet the requirement $\mathbf{E}[W_2] \leqslant \overline{W}$. If $\overline{W} = w_{FCFS}$, only FCFS can meet the requirement and no revenue can be made.
2. $\overline{W} < \frac{w_{FCFS}}{1-\rho}$. Otherwise the requirement $\mathbf{E}[W_2] \leqslant \overline{W}$ is fulfilled even when all customers go to queue 1 under strict priority. In this case, strict priority maximizes the revenue.

4 When and How much does Hybrid help?

In this section, we solve the constrained optimization problem (5) to investigate when and how much Hybrid (or partial priority) helps. The main goal of this section is to prove Theorems 1 and 2, which are stated below for easy reference, but will be proved later in the section. These theorems use notation which is explained in Table 1 and will be defined in this section.

Notably, a straightforward corollary of Theorem 1, Corollary 1, shows that partial priority increases revenue only if *both* the $\bar{\$}$ and $\overline{W}$ restrictions exist.

Theorem 1 (When?)**.** *Hybrid (or partial priority) increases the revenue compared with strict priority if and only if*

$$\frac{\overline{W} - w_{FCFS}}{\overline{W}} < \rho \bar{F}_C \left(\frac{\bar{\$}}{\overline{W} \cdot \rho} \right). \tag{6}$$

Theorem 2 (How much?)**.** *If the condition (6) in Theorem 1 holds, then*

$$\textit{Improvement Ratio} = \frac{\lambda_1^*}{\widehat{\lambda_1}},$$

where λ_1^* is the unique solution of

$$\frac{\lambda}{\lambda_1^*} \cdot g\left(\frac{\lambda_1^*}{\lambda} \right) = \frac{\bar{\$}}{\overline{W} - w_{FCFS}},$$

and g is defined in Definition 1.

We proceed as follows: In Section 4.1 we introduce some notation to simplify the optimization problem (5) from Section 3. Then in Section 4.2, we give the proof of Theorem 1 and Theorem 2. Finally in Section 4.3, we discuss and present analogous theorems for a variant of our model in which λ_1 cannot be limited by the service provider. Table 1 summarizes new notation introduced in this section.

4.1 Simplification of the optimization problem

The goal of this section is to rewrite the optimization problem (5). To do this, we introduce some new notation. We first define a function $g(\cdot)$ to be the inverse of $\bar{F}_C(\cdot)$. Under the assumption that $\bar{F}_C(\cdot)$ is invertible and continuous, we define the continuous function g as follows.

Table 1: Additional Notation Table

Notation	Mathematical Definition	Meaning
$g(x)$	See Definition 1	Inverse of $\bar{F}_C(\cdot)$
$\theta(\lambda_1, q)$	See Definition 2	The maximum price charged without the price cap
$\widehat{\lambda_1}$	$\mu\left(1 - \frac{w_{FCFS}}{W}\right)$	The maximum arrival rate under strict priority to ensure $\mathbf{E}[W_2] \leqslant W$
superscript * (e.g. λ_1^*)	-	Values in the optimal solution
superscript $'$ (e.g. λ_1')	-	Temporary notation for proofs
Improvement Ratio	See Definition 4	The ratio of the optimal revenue under Hybrid to that under strict priority

Definition 1 ($g(x)$)**.** *For any $x \in [0,1]$ define $g(x)$ to be*

$$g(x) := \bar{F}_C^{-1}(x).$$

We next define a shorthand $\theta(\lambda_1, q)$. For now the definition seems arbitrary, but we will see that this term emerges in the proof of Lemma 1. Intuitively, for given λ_1 and q, we can think of $\theta(\lambda_1, q)$ as the maximum price that the service provider can charge to ensure that at least $\frac{\lambda_1}{\lambda}$ fraction of the customers are willing to buy priority. In other words, $\theta(\lambda_1, q)$ is an upper bound on the price that the service provider can charge, given λ_1 and q, and given no price cap.

Definition 2 ($\theta(\lambda_1, q)$)**.** *Define*

$$\theta(\lambda_1, q) := g\left(\frac{\lambda_1}{\lambda}\right) \cdot \frac{\lambda}{\lambda_1}\left(\mathbf{E}[W_2 \mid \lambda_1, q] - w_{FCFS}\right).$$

Now we give the simplification of the optimization problem (5).

Lemma 1 (Simplification of the optimization problem)**.** *The optimization problem (5) is equivalent to the optimization problem in (7):*

$$\boxed{\begin{array}{ll} \underset{\lambda_1, q}{\textit{maximize}} & \$ \cdot \lambda_1 \\ \textit{s.t.} & \$ = \min\{\theta(\lambda_1, q), \bar{\$}\}, \\ & \mathbf{E}[W_2] \leqslant \overline{W}, \\ & 1 \geqslant q \geqslant 0. \end{array}} \tag{7}$$

Proof. First, we simplify the constraint (4). By the conservation law (see [37] or proof of Proposition 6),

$$\frac{\lambda_1}{\lambda} \cdot \mathbf{E}[W_1] + \frac{\lambda_2}{\lambda} \cdot \mathbf{E}[W_2] = w_{FCFS}.$$

Thus we have that

$$\begin{aligned} \mathbf{E}[W_2] - \mathbf{E}[W_1] &= \frac{\lambda}{\lambda_1}\left(\mathbf{E}[W_2] - \frac{\lambda_1}{\lambda} \cdot \mathbf{E}[W_1] - \frac{\lambda_2}{\lambda} \cdot \mathbf{E}[W_2]\right) \\ &= \frac{\lambda}{\lambda_1}\left(\mathbf{E}[W_2] - w_{FCFS}\right). \end{aligned}$$

Thus the constraint (4) can be reformulated into

$$\frac{\lambda_1}{\lambda} \leqslant \bar{F}_C\left(\frac{\$}{\mathbf{E}[W_2] - \mathbf{E}[W_1]}\right) \iff \frac{\lambda_1}{\lambda} \leqslant \bar{F}_C\left(\frac{\lambda_1}{\lambda} \cdot \frac{\$}{\mathbf{E}[W_2] - w_{FCFS}}\right).$$

We now apply g on both sides. Noticing that g is a decreasing function, we have that the constraint (4) is equivalent to

$$g\left(\frac{\lambda_1}{\lambda}\right) \geqslant \frac{\lambda_1}{\lambda} \cdot \frac{\$}{\mathbf{E}[W_2] - w_{FCFS}},$$

which is further equivalent to

$$\$ \leqslant \theta(\lambda_1, q).$$

Observe that the only two constraints on \$ are $\$ \leqslant \bar{\$}$ and $\$ \leqslant \theta(\lambda_1, q)$. Hence we know that, given λ_1, q, the revenue maximizing \$ should be at most $\min\{\theta(\lambda_1, q), \bar{\$}\}$. But this upper bound is actually an equality because as long as λ_1 is fixed, higher price is better. Hence we have that $\$ = \min\{\theta(\lambda_1, q), \bar{\$}\}$ in (7). This finishes the proof. □

4.2 When and how much does Hybrid (partial priority) help?

In this section, we will solve the optimization problem (7) to prove the main theorems. Since the feasible set of this optimization problem is compact and the optimization is finite, an optimal solution exists. Throughout this paper, we will use the superscipt * to denote the optimal solution. Specifically, the triple $(\lambda_1^*, q^*, \$^*)$ denotes the optimal values of each of the decision variables. Furthermore $\mathbf{E}[W_2^*] := \mathbf{E}[W_2 \mid \lambda_1^*, q^*]$ and $\theta^* := \theta(\lambda_1^*, q^*)$.

Lemma 2 provides some conditions on the optimal solution.

Lemma 2. *At least one of the following conditions must hold:*

1. $\mathbf{E}[W_2^*] = \overline{W}$ *and* $\theta^* = \bar{\$}$.
2. $q^* = 1$.

The first condition says that both the constraints on $\overline{W}$ and on $\bar{\$}$ are binding. The second condition says that strict priority is optimal.

Proof. We prove this by contradiction. Suppose both items (1) and (2) above are false. Then we have $q^* < 1$. We discuss different cases and show that all of them lead to contradictions.

We perturb λ_1^* and q^* in all cases to prove contradictions. This approach is feasible because $\lambda_1^* \in (0, \lambda)$ and $q^* \in (0, 1)$. To justify this, note that if $\lambda_1^* = 0$, the revenue would be zero which is clearly suboptimal. If $\lambda_1^* = \lambda$, then $\theta^* = 0$ because $g(1) =$

0, which also leads to zero revenue. If $q^* = 0$, then $\theta(\lambda_1^*, q^* = 0) < 0$ because $\mathbf{E}[W_2 \mid \lambda_1^*, q^* = 0] < w_{FCFS}$. Finally, if $q^* = 1$, condition (2) above holds.

Case 1: $\mathbf{E}[W_2^*] < \overline{W}$. We increase λ_1^* to $\lambda_1' = \lambda_1^* + \epsilon$. Now we perturb q^* to q' such that $\theta(\lambda_1', q') = \theta^*$. This can be done when ϵ is small enough by the continuity of θ. This change of value can still guarantee that $\mathbf{E}[W_2'] \leqslant \overline{W}$ if ϵ is small enough. In this way, λ_1 is increased and the price is the same, which increases the revenue.

Case 2: $\mathbf{E}[W_2^*] = \overline{W}$ and $\theta^* > \bar{\$}$. In this case, the revenue is $\bar{\$} \cdot \lambda_1^*$. Then there exists a small enough $\epsilon > 0$ such that we can increase λ_1^* to $\lambda_1' = \lambda_1^* + \epsilon$ and decrease q^* to q' making $\mathbf{E}[W_2 \mid \lambda_1', q'] = \mathbf{E}[W_2^*]$. We now make ϵ small enough so that $\theta(\lambda_1', q') \geqslant \bar{\$}$ still holds. In this way, λ_1 is increased, and the price is still $\bar{\$}$. This leads to a larger revenue which is a contradiction.

Case 3: $\mathbf{E}[W_2^*] = \overline{W}, \theta^* < \bar{\$}$. In this case, the revenue is

$$\text{Revenue}^* = \theta^* \cdot \lambda_1^* = \lambda \cdot g\left(\frac{\lambda_1^*}{\lambda}\right) \cdot (\mathbf{E}[W_2] - w_{FCFS}).$$

Thus we can decrease λ_1^* to $\lambda_1' = \lambda_1^* - \epsilon$ and increase q^* to q' such that $\mathbf{E}[W_2 \mid \lambda_1', q'] = \mathbf{E}[W_2^*]$. We now make ϵ small enough such that $\theta(\lambda_1', q') \leqslant \bar{\$}$ still holds. In this way, the revenue is also increased because g is decreasing with respect to λ_1, which is again a contradiction.

Combining all those three cases yields the proof. □

We define the notation $\widehat{\lambda_1}$ for the following proofs. Intuitively, $\widehat{\lambda_1}$ is the maximum arrival rate under strict priority to ensure $\mathbf{E}[W_2] \leqslant \overline{W}$. By using (18), which says $\mathbf{E}[W_2 \mid \lambda_1, q = 1] = \frac{w_{FCFS}}{1-\rho_1}$, we can further get a closed form for $\widehat{\lambda_1}$.

Definition 3 ($\widehat{\lambda_1}$)**.** *Define $\widehat{\lambda_1}$ to be the solution of $\mathbf{E}\left[W_2 \mid \widehat{\lambda_1}, q = 1\right] = \overline{W}$. Mathematically,*

$$\widehat{\lambda_1} = \mu\left(1 - \frac{w_{FCFS}}{\overline{W}}\right). \tag{8}$$

Now we can prove our main lemma characterizing when Hybrid helps increase the revenue.

Lemma 3. *Hybrid (or partial priority) helps increase the revenue compared with strict priority if and only if*

$$\theta(\widehat{\lambda_1}, q = 1) > \bar{\$}. \tag{9}$$

Proof. We first prove that $\theta(\widehat{\lambda_1}, q = 1) > \bar{\$}$ is a *necessary condition* for Hybrid to help.

Suppose by contradiction that $\theta(\widehat{\lambda_1}, q = 1) \leqslant \bar{\$}$ and assume that Hybrid still helps. This is saying that we obtain the optimal revenue with $q^* < 1$. By Lemma 2, we have that $\mathbf{E}[W_2^*] = \overline{W}$ and $\theta^* = \bar{\$}$. Then

$$\text{Revenue}^* = \theta^* \cdot \lambda_1^* = \lambda \cdot g\left(\frac{\lambda_1^*}{\lambda}\right) \cdot (\mathbf{E}[W_2^*] - w_{FCFS}) = \lambda \cdot g\left(\frac{\lambda_1^*}{\lambda}\right) \cdot \left(\overline{W} - w_{FCFS}\right).$$

On the other hand, if we set $\lambda_1 = \widehat{\lambda_1}$ under strict priority, the revenue obtained is

$$\text{Revenue}' = \theta(\widehat{\lambda_1}, q = 1) \cdot \widehat{\lambda_1} = \lambda \cdot g\left(\frac{\widehat{\lambda_1}}{\lambda}\right) \cdot \left(\overline{W} - w_{FCFS}\right).$$

Since the optimal $q^* < 1$, we know that $\text{Revenue}^* > \text{Revenue}'$, which indicates $\lambda_1^* < \widehat{\lambda_1}$ because g is decreasing.

But this leads to the contradiction:

$$\overline{W} = \mathbf{E}[W_2 \mid \lambda_1^*, q^*] < \mathbf{E}[W_2 \mid \lambda_1^*, q = 1] \leqslant \mathbf{E}\left[W_2 \mid \widehat{\lambda_1}, q = 1\right] = \overline{W}.$$

We next prove that $\theta(\widehat{\lambda_1}, q = 1) > \bar{\$}$ is a *sufficient condition* for Hybrid to help. In this case, the optimal revenue under strict priority is achieved at $\widehat{\lambda_1}$ because any parameters satisfying the restriction under strict priority satisfy $\lambda_1 \leqslant \widehat{\lambda_1}$ and $\$ \leqslant \bar{\$}$.

Now we can pick a small $\epsilon > 0$ and set $q' = 1 - \epsilon$. Accordingly we can increase $\widehat{\lambda_1}$ to λ_1' such that $\mathbf{E}\left[W_2 \mid \widehat{\lambda_1}, q = 1\right] = \mathbf{E}[W_2 \mid \lambda_1', q']$. Pick ϵ small enough such that

$\theta(\lambda_1', q') \geqslant \bar{\$}$ still holds. In this way, the revenue is $\bar{\$} \cdot \lambda_1' > \bar{\$} \cdot \widehat{\lambda_1}$ which is the optimal revenue under strict priority. □

Our main theorem characterizing the necessary and sufficient condition for Hybrid (or partial priority) to help follows directly by simplifying the condition in Lemma 3.

Theorem 1. *Hybrid (or partial priority) helps increase the revenue compared with strict priority if and only if*

$$\frac{\overline{W} - w_{FCFS}}{\overline{W}} < \rho \bar{F}_C \left(\frac{\bar{\$}}{\overline{W} \cdot \rho} \right). \tag{10}$$

Proof. We substitute (8) into condition (9), yielding:

$$\theta(\widehat{\lambda_1}, q = 1) = g\left(\frac{\widehat{\lambda_1}}{\lambda}\right) \cdot \frac{\lambda}{\widehat{\lambda_1}} \left(\mathbf{E}\left[W_2 \mid \widehat{\lambda_1}, q = 1 \right] - w_{FCFS} \right) = g\left(\frac{\widehat{\lambda_1}}{\lambda}\right) \cdot \frac{\lambda}{\widehat{\lambda_1}} \left(\overline{W} - w_{FCFS} \right). \tag{11}$$

Thus condition (9) is equivalent to:

$$\begin{aligned}
\theta(\widehat{\lambda_1}, q = 1) > \bar{\$} &\iff g\left(\frac{\widehat{\lambda_1}}{\lambda}\right) > \frac{\widehat{\lambda_1}}{\lambda} \cdot \frac{\bar{\$}}{\overline{W} - w_{FCFS}} \qquad \text{by (11)} \\
&\iff g\left(\frac{\widehat{\lambda_1}}{\lambda}\right) > \frac{1}{\lambda} \cdot \mu \cdot \frac{\overline{W} - w_{FCFS}}{\overline{W}} \cdot \frac{\bar{\$}}{\overline{W} - w_{FCFS}} \\
&\iff g\left(\frac{\widehat{\lambda_1}}{\lambda}\right) > \frac{\bar{\$}}{\overline{W} \cdot \rho} \\
&\iff \frac{\widehat{\lambda_1}}{\lambda} < \bar{F}_C \left(\frac{\bar{\$}}{\overline{W} \cdot \rho} \right) \\
&\iff \frac{\mu(1 - \frac{w_{FCFS}}{W})}{\lambda} < \bar{F}_C \left(\frac{\bar{\$}}{\overline{W} \cdot \rho} \right) \\
&\iff \frac{\overline{W} - w_{FCFS}}{\overline{W}} < \rho \bar{F}_C \left(\frac{\bar{\$}}{\overline{W} \cdot \rho} \right).
\end{aligned}$$

□

A straightforward corollary of Theorem 1 is presented below.

Corollary 1. *Partial priority increases revenue only if* both *the* $\bar{\$}$ *and* $\overline{W}$ *restrictions exist. In other words, partial priority policies do not increase the revenue when* $\bar{\$} \to \infty$ *or* $\overline{W} \to \infty$.

Proof. We only need to examine whether condition (10) holds when $\bar{\$} \to \infty$ or $\overline{W} \to \infty$.

When $\bar{\$} \to \infty$: The left hand side of (10) is positive and the right hand side goes to 0.

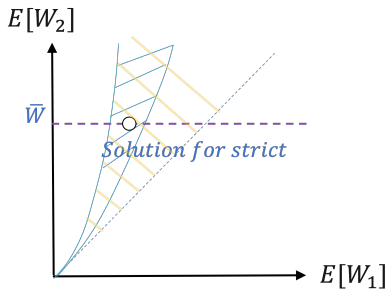
When $\overline{W} \to \infty$: The right hand side of (10) is smaller than 1 since both $\rho < 1$ and $\bar{F}_C(\cdot) \leqslant 1$. The left hand side goes to 1. □

Remark 1. *We present some intuitions for why* both *restrictions are necessary for partial priority to improve revenue over strict priority.*

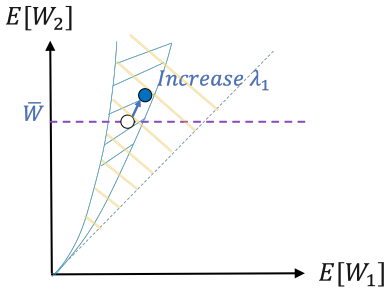
Under strict priority, because of the $\overline{W}$ *restriction, the service provider must limit the rate of sales of priority passes. The few customers who buy priority have high time-sensitivity (tail of* C*), so they are willing to pay a lot, but the service provider is also limited by the* $\bar{\$}$ *restriction. This leaves money on the table (the service provider does not get to make as much revenue as they would like). By applying partial priority, the class 2 customers experience less waiting. Hence the service provider is allowed to sell more priority passes (at the same price), while still adhering to the* $\overline{W}$ *restriction, thus making more revenue.*

Pictorially, let's revisit Figure 1. Since class 1 customers have (partial) priority over class 2 customers, we only show the half plane where $\mathbf{E}[W_1] < \mathbf{E}[W_2]$. *Suppose the optimal strict priority policy yields the waiting time pair shown in Figure 3(a) and suppose that the optimal price already meets the constraint* $\bar{\$}$. *In order to increase the revenue, one wants to increase* λ_1*; However that yields an infeasible waiting time pair under strict priority (as shown in Figure 3(b)). The way that Hybrid can bring us more money is shown in Figure 3(c): One can use Hybrid to make the solution feasible*

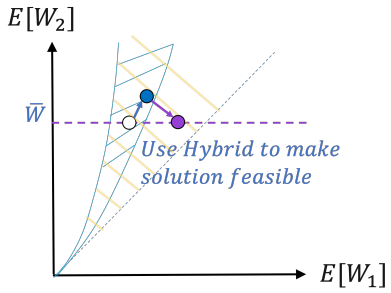
again while keeping the λ_1 and the price the same as those in Figure 3. The question of whether Hybrid (and partial priorities) increases revenue depends on whether or not customers are still willing to pay $\bar{\$}$ after transitioning to the new waiting time pair shown in Figure 3(c).



(a) The expected waiting time pair under the optimal strict priority policy.



(b) Increasing λ_1 will increase the revenue. However, the solution now violates the restriction $\bar{W}$.



(c) Using Hybrid to make the solution feasible again. The resulting waiting time pair only lies in the achievability region of Hybrid but not strict priority.

Fig. 3: Pictorial illustration for how partial priority may improve the revenue.

We see that both restrictions are important in this intuition, and without any one of them, the intuition breaks.

Finally, we present a theorem that demonstrates the extent to which Hybrid outperforms strict priority. The metrics we use are the *improvement ratio* and the *improvement amount*.

Definition 4. *Let Revenue*$(q = 1)$ *denote the optimal revenue under strict priority and Revenue** *denote the optimal revenue under Hybrid (or partial priority). Define the improvement ratio and the improvement amount to be*

$$\textit{Improvement Ratio} := \frac{\textit{Revenue}^*}{\textit{Revenue}(q=1)}.$$

$$\textit{Improvement Amount} := \textit{Revenue}^* - \textit{Revenue}(q=1).$$

Now we prove our main theorem characterizing the improvement ratio and the improvement amount of Hybrid.

Theorem 2 (How much?)**.** *If condition (10) in Theorem 1 holds, then*

$$\textit{Improvement Ratio} = \frac{\lambda_1^*}{\widehat{\lambda_1}}, \quad \textit{and} \quad \textit{Improvement Amount} = \left(\frac{\lambda_1^*}{\widehat{\lambda_1}} - 1\right) \cdot \widehat{\lambda_1} \cdot \bar{\$},$$

where λ_1^* *is the unique solution of*

$$\frac{\lambda}{\lambda_1^*} \cdot g\left(\frac{\lambda_1^*}{\lambda}\right) = \frac{\bar{\$}}{\overline{W} - w_{FCFS}}. \tag{12}$$

Proof. Condition (10) in Theorem 1 is equivalent to condition (9) in Lemma 3, which is $\theta(\widehat{\lambda_1}, q = 1) > \bar{\$}$. In this case, the optimal revenue under strict priority is $\widehat{\lambda_1} \cdot \bar{\$}$, which is achieved when $\lambda_1 = \widehat{\lambda_1}$. The reason why no larger revenue can be achieved is that under strict priority, $\lambda_1 \leqslant \widehat{\lambda_1}$ because of the $\overline{W}$ restriction and $\$ \leqslant \bar{\$}$. This gives that Revenue$(q = 1) = \widehat{\lambda_1} \cdot \bar{\$}$.

On the other hand, since condition (10) holds, by Theorem 1, we know that the optimal $q^* < 1$. Thus by Lemma 2 we know that

$$\theta^* = \bar{\$}, \quad \mathbf{E}\left[W_2^*\right] = \overline{W}. \tag{13}$$

This shows that the improvement ratio is given by:

$$\text{Improvement Ratio} = \frac{\bar{\$} \cdot \lambda_1^*}{\bar{\$} \cdot \widehat{\lambda_1}} = \frac{\lambda_1^*}{\widehat{\lambda_1}}.$$

The expression for λ_1^* can be solved from the definition of $\theta(\lambda_1, q)$ and (13), which is equivalent to

$$\frac{\lambda}{\lambda_1^*} \cdot g\left(\frac{\lambda_1^*}{\lambda}\right) = \frac{\bar{\$}}{\overline{W} - w_{FCFS}}.$$

Note that the left hand side of the above equation is continuous, monotonic with λ_1^*, and ranges from ∞ to 0 when λ_1^* takes on values from 0 to λ. Thus there exists a unique solution for λ_1^*.

Finally, the formula for the improvement amount is given by

$$\text{Improvement Amount} = (\text{Improvement Ratio} - 1) \cdot \text{Revenue}(q = 1).$$

□

An immediate corollary of Theorem 2 is that the improvement ratio may go to infinity under extreme conditions on $\bar{\$}$ and $\overline{W}$.

Corollary 2 (Infinite improvement ratio)**.** *As* $\overline{W} \to w_{FCFS}$ *from above and* $\frac{\bar{\$}}{\overline{W} - w_{FCFS}} \to 0$ *from above, the improvement ratio goes to infinity.*

Proof. By equation (8), we know that $\widehat{\lambda_1} \to 0$ from above. By equation (12), we know that $g\left(\frac{\lambda_1^*}{\lambda}\right) \to 0$ from above, which indicates that $\lambda_1^* \to \lambda$ from below. Thus by

Theorem 2, we have:

$$\text{Improvement Ratio} = \frac{\lambda_1^*}{\widehat{\lambda_1}} \to \infty.$$

□

4.3 Variant model: When the service provider can't limit priority ticket sales

In this section, we briefly explore a variant of our model where the service provider is unable to limit the number of priority passes sold. In other words, whenever a customer wants to buy priority, she can buy it. As stated in Section 3 (see the discussion after (4)), this implies that inequality (4) must become an equality, transforming the optimization problem into the following form:

$$\begin{aligned} \underset{\lambda_1, q, \$}{\text{maximize}} \quad & \$ \cdot \lambda_1 \\ \text{s.t.} \quad & \frac{\lambda_1}{\lambda} = \bar{F}_C\left(\frac{\$}{\mathbf{E}[W_2] - \mathbf{E}[W_1]}\right), \\ & \mathbf{E}[W_2] \leqslant \overline{W}, \\ & \$ \leqslant \bar{\$}, \\ & 0 \leqslant q \leqslant 1. \end{aligned} \tag{14}$$

Note that under strict priority, it is now possible that no feasible solution exists. For instance, if the price cap is low, a large number of customers may want to purchase priority because of the low price. Since all customers desiring priority are able to buy it, there will be too many customers in the priority queue, which may make it impossible to ensure that $\mathbf{E}[W_2] \leqslant \overline{W}$. By contrast, a solution is always attainable in the Hybrid (or partial priority) model, as we can make $\mathbf{E}[W_1]$ and $\mathbf{E}[W_2]$ closer to make priority less attractive.

Theorems similar to those in our original model also hold in this variation. The proofs follow a similar structure (see Appendix C), and we omit them here for brevity.

Theorem 3. *Hybrid (or partial priority) helps if and only if*

$$\frac{\overline{W} - w_{FCFS}}{\overline{W}} < \rho \bar{F}_C \left(\frac{\bar{\$}}{\overline{W} \cdot \rho} \right). \tag{15}$$

Proof. See Appendix C. □

Theorem 4. *Assume the condition (15) in Theorem 3 holds.*

1. *If there is no* $\lambda_1 \leqslant \widehat{\lambda_1}$ *satisfying* $\theta(\lambda_1, 1) = \bar{\$}$*, then there is no feasible solution under strict priority (which means Hybrid or partial priority beats strict priority since there always exists a solution under Hybrid);*
2. *Otherwise, let* $\lambda_1^{q=1} := \max_{\lambda_1} \left\{ \theta(\lambda_1, 1) = \bar{\$} \mid \lambda_1 \leqslant \widehat{\lambda_1} \right\}$*. Then we have that*

$$\textit{Improvement Ratio} = \frac{\lambda_1^*}{\lambda_1^{q=1}}, \quad \textit{and} \quad \textit{Improvement Amount} = \left(\frac{\lambda_1^*}{\lambda_1^{q=1}} - 1 \right) \cdot \lambda_1^{q=1} \bar{\$},$$

where λ_1^* *is the unique solution of*

$$\frac{\lambda}{\lambda_1^*} \cdot g\left(\frac{\lambda_1^*}{\lambda} \right) = \frac{\bar{\$}}{\overline{W} - w_{FCFS}}.$$

Proof. See Appendix C. □

Importantly: the condition under which Hybrid improves revenue is the same in both models (compare Theorem 1 with Theorem 3), but the improvement is greater in this variant than in the original model (compare Theorem 2 with Theorem 4). The latter point follows directly from the fact that $\lambda_1^{q=1} \leqslant \widehat{\lambda_1}$ by definition.

5 Numerical Study with Pareto Disutility

To illustrate the improvement possible under partial priority, we evaluate our results in Section 4 under the realistic assumption that C is distributed as a Pareto distribution. Intuitively, a customer's impatience factor C (the monetary cost a customer experiences per unit waiting time) is likely related to the customer's wealth, which has long been known to often follow a Pareto distribution ([38, 39].) Mathematically, we assume a Pareto type II distribution with tail parameter α where

$$\bar{F}_C(x) = \left(\frac{1}{1+x}\right)^{\alpha}, \quad x \geqslant 0. \tag{16}$$

As we saw in in Corollary 2, the improvement ratio can approach infinity under some $\overline{\$}$ and $\overline{W}$. In this section, we examine the improvement under realistic parameters.

5.1 Disney World setting

The parameters that need to be decided are as follows: the parameter α for the Pareto distribution, the service time distribution S, the total traffic/load $\rho = \frac{\lambda}{\mu}$ and the values of the two restrictions: $\overline{\$}$ and $\overline{W}$.

In this section, we consider a particular setting of the above parameters which is reasonably consistent with Disney World, and look at the improvement ratio under these parameters. In Section 5.2, we will explore a range of parameter values and look at the performance for that range.

Pareto parameter α: We set $\alpha = 1.5$ to satisfy the 80-20 rule (also known as Pareto Principle, saying that 80% of the wealth is owned by 20% of the population, [38]). Under this distribution, on average people are willing to pay about \$ 1.41 per minute waiting time, which seems reasonable.

Service time distribution S: In the case of Disney, the service time S can be approximated by a Deterministic distribution with value 1 minute.

Total traffic/load ρ: We set the total load $\rho = 0.98$. Intuitively, ρ is the fraction of time that the server is busy. In Disney World, it is nearly impossible to see the server idle, given that single-pass lightning passes are only sold for the most in-demand attractions ([22]). Under this value of load, if all customers are served in FCFS order, the average waiting time will be about 25 minutes, which is also reasonable.

Outer restrictions $\overline{\$}$ and $\overline{W}$: We restrict the mean waiting time for class 2 customers to be no more than $\overline{W} = 30$ minutes. We set the price cap to be $\overline{\$} = 25$ dollars per pass per person. The price cap is consistent with what we observe in reality (see "Lightning Lane Single Pass Pricing at Disney World" in [40]).

Under the above parameters, Theorem 2 shows that the improvement ratio is about *1.53*, which means Hybrid increases the revenue by more than 50 percent. The improvement amount is about *2.43*, which can be interpreted as 2.43 dollars per minute for the attraction (about 1200 dollars per day). To have a better understanding, these numbers mean that Hybrid will increase the number of customers buying a lightning pass for this attraction from about 80 per day to about 120 per day. This in turn will translate to an increase in daily revenue from selling lightning passes for the attraction from 2400 dollars to 3600.

5.2 Exploring a range of parameter settings

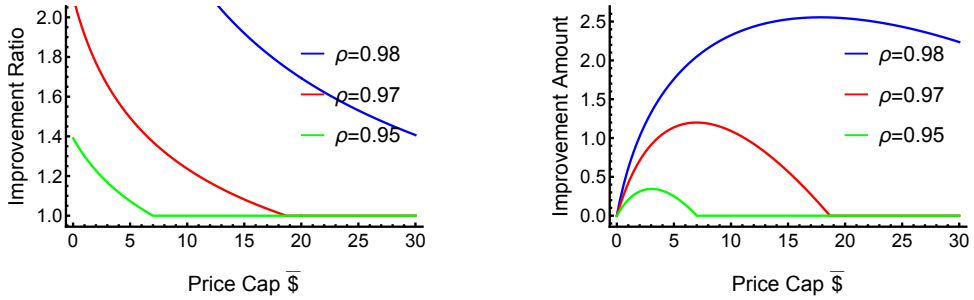
To get further insights, we plot the improvement ratio and the improvement amount from Theorem 2 under more parameter combinations.

Effect of constraints on $\overline{\$}$ and $\overline{W}$: Figure 4 and Figure 5, respectively, show how the improvement ratio is affected by the constraints on $\overline{\$}$ and $\overline{W}$. All other parameters are set to be the same as those in Section 5.1. As shown in the figures, *the improvement ratio is higher when the restrictions become tighter* (i.e., $\overline{\$}$ and/or $\overline{W}$ become smaller). This is intuitive as when either $\overline{\$}$ or $\overline{W}$ is unconstrained, strict priority is optimal. On the other hand, while the improvement ratio is monotonic, the improvement amount

is not. This is because when the restrictions are extremely tight, the service provider can hardly make money.

Another observation is that *the improvement ratio is higher under high traffic.* This is also intuitive because when the traffic is low, the regular line customers are not suffering from a long waiting time, thus we do not need Hybrid to reduce the regular line waiting time.

Notice that the improvement ratio diverges at some limit in Figure 5. This limit is $\overline{W} \to w_{FCFS}$. At this limit, $\widehat{\lambda_1}$ in Theorem 2 goes to 0, accounting for the diverging improvement ratio.



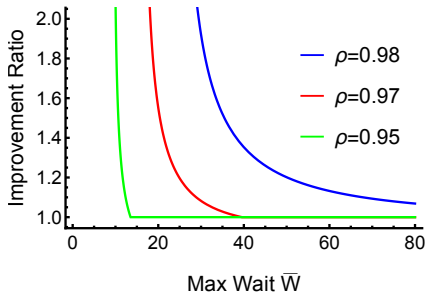
(a) The improvement ratio.

(b) The improvement amount.

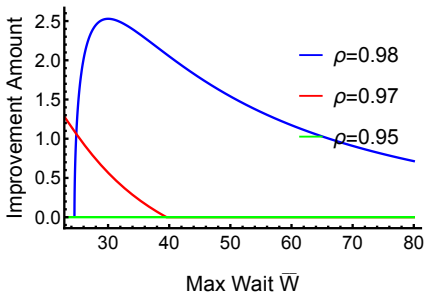
Fig. 4: Hybrid (or partial priority) generates significant improvement under high traffic (e.g., $\rho = 0.98$.) Also, the improvement ratio grows as \$ becomes smaller, but the improvement amount is not monotone. In this set of experiments, we set $\alpha = 1.5$, $S = 1$ (deterministic) and $\overline{W} = 30$.

Effect of Service Time Distributions: In Section 5.1, we assumed that the service time distribution is deterministic. We explore the impact of higher variance in service time in Figure 6.

As shown in Figure 6, *Hybrid (or partial priority) helps increase the revenue at lower load when the service time distribution has higher variance.* The intuition is as follows: With a higher variance in service time, the average waiting time for customers increases, which makes the restriction $\overline{W}$ relatively more restrictive. Consequently a



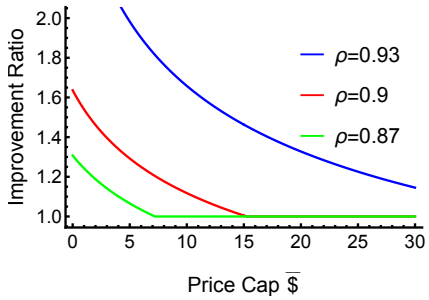
(a) The improvement ratio.



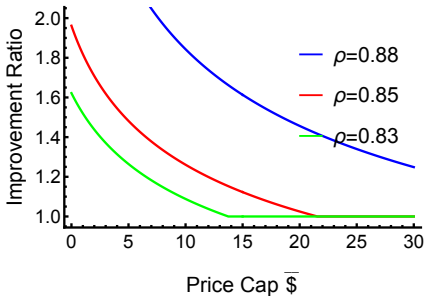
(b) The improvement amount.

Fig. 5: Hybrid (or partial priority) generates significant improvement under high traffic (e.g., $\rho = 0.98$.) Also, the improvement ratio grows as $\overline{W}$ becomes smaller, but the improvement amount is not monotone. In this set of experiments, we set $\alpha = 1.5$, $S = 1$ (deterministic) and $\overline{\$} = 15$.

smaller load is required to achieve the same improvement ratio when service time variance is higher.



(a) Variance of service time is 2.



(b) Variance of service time is 5.

Fig. 6: As the variance of service time becomes higher, Hybrid (or partial priority) helps increase the revenue at lower load. In this set of experiments, we set $\alpha = 1.5$ and $\overline{W} = 30$. In (a) S follows any distribution with mean 1 and variance 2. In (b) S follows any distribution with mean 1 and variance 5.

The Pareto parameter α: We also evaluated the improvement ratio when the customer disutility follows a more heavy-tail Pareto distribution ($\alpha = 1.1$ instead of $\alpha = 1.5$). It turns out that the value of α does not affect the improvement ratio significantly (see Figure 9 in Appendix D compared with Figure 4).

6 Conclusion

We consider the setting where a service provider wishes to increase revenue by leveraging the fact that time-sensitive customers are willing to pay for shorter waiting times. This paper studies whether partial priority can bring in more revenue than strict priority.

The main insight of this paper is that, despite the flexibility of partial priority, it only increases revenue if there are "tight" restrictions on the service provider (see Corollary 1 and Remark 1). Specifically we need restrictions on *both* the maximum mean waiting time $\overline{W}$ and the maximum price, $\bar{\$}$. Such waiting time and price restrictions, however, are common in practice, so there are in fact situations where partial priority is helpful.

We provide a necessary and sufficient condition on the tightness of the restrictions needed for partial priority to increase revenue (Theorem 1). We also provide an analytical characterization of the improvement ratio and amount of Hybrid over strict priority (Theorem 2).

The key steps in our work are Lemmas 2 and 3. Lemma 2 uses perturbation analysis to create a characterization of the optimal solution. Lemma 3 builds upon this characterization to distill a necessary and sufficient condition for partial priority to help improve the revenue. Theorem 1 then simplifies the condition in Lemma 3 into closed-form explicit expressions.

We close this paper by discussing directions for future research. While our current work focuses on a two-queue system, a natural extension would be to consider systems with multiple queues and multiple priority levels. Although our Hybrid policy can be easily generalized by partitioning probabilities across the multiple queues, it is not clear how the restrictions should generalize. For example, it is at this point unclear whether a single price cap is sufficient for partial priority to increase the revenue, or if different price caps are needed for each priority level. Additionally, exploring scenarios

where the total arrival rate λ is variable and depends on the pricing mechanism is complementary to our model and left for future work.

References

[1] Hassin, R.: Decentralized regulation of a queue. Management Science **41**(1), 163–173 (1995)

[2] Hsu, V.N., Xu, S.H., Jukic, B.: Optimal scheduling and incentive compatible pricing for a service system with quality of service guarantees. Manufacturing & Service Operations Management **11**(3), 375–396 (2009)

[3] Haviv, M., Winter, E.: An optimal mechanism charging for priority in a queue. Operations Research Letters **48**(3), 304–308 (2020)

[4] Hassin, R., Haviv, M.: To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems vol. 59. Springer, New York (2003)

[5] Mendelson, H., Whang, S.: Optimal incentive-compatible priority pricing for the M/M/1 queue. Operations Research **38**(5), 870–883 (1990)

[6] Alperstein, H.: Note—optimal pricing policy for the service facility offering a set of priority prices. Management Science **34**(5), 666–671 (1988)

[7] Coffman Jr, E.G., Mitrani, I.: A characterization of waiting time performance realizable by single-server queues. Operations Research **28**(3-part-ii), 810–821 (1980)

[8] Federgruen, A., Groenevelt, H.: Characterization and optimization of achievable performance in general queueing systems. Operations Research **36**(5), 733–741 (1988)

[9] Lederer, P.J., Li, L.: Pricing, production, scheduling, and delivery-time competition. Operations research **45**(3), 407–420 (1997)

[10] Afeche, P., Baron, O., Milner, J., Roet-Green, R.: Pricing and prioritizing time-sensitive customers with heterogeneous demand rates. Operations Research **67**(4), 1184–1208 (2019)

[11] Katta, A.-K., Sethuraman, J.: Pricing strategies and service differentiation in queues—a profit maximization perspective. Department of Industrial Engineering and Operations Research, Columbia University (2005)

[12] Maglaras, C., Yao, J., Zeevi, A.: Optimal price and delay differentiation in large-scale queueing systems. Management science **64**(5), 2427–2444 (2018)

[13] Gurvich, I., Lariviere, M.A., Ozkan, C.: Coverage, coarseness, and classification: Determinants of social efficiency in priority queues. Management Science **65**(3), 1061–1075 (2019)

[14] Afeche, P., Pavlin, J.M.: Optimal price/lead-time menus for queues with customer choice: Segmentation, pooling, and strategic delay. Management Science **62**(8), 2412–2436 (2016)

[15] Afeche, P.: Incentive-compatible revenue management in queueing systems: Optimal strategic delay. Manufacturing & Service Operations Management **15**(3), 423–443 (2013)

[16] Hassin, R., Haviv, M.: Who should be given priority in a queue? Operations Research Letters **34**(2), 191–198 (2006)

[17] Haviv, M., van der Wal, J.: Waiting times in queues with relative priorities. Operations Research Letters **35**(5), 591–594 (2007)

[18] Hassin, R., Puerto, J., Fernández, F.R.: The use of relative priorities in optimizing the performance of a queueing system. European Journal of Operational Research **193**(2), 476–483 (2009)

[19] Haviv, M., Ravner, L.: Strategic bidding in an accumulating priority queue: equilibrium analysis. Annals of Operations Research **244**(2), 505–523 (2016)

[20] Moshe, S., Oz, B.: Charging more for priority via two-part tariff for accumulating priorities. European Journal of Operational Research **304**(2), 652–660 (2023)

[21] Disney Lightning Lane Passes. https://disneyworld.disney.go.com/lightning-lane-passes/

[22] Overview of Disney World Lightning Lane Passes. https://www.undercovertourist.com/blog/disney-world-lightning-lane/

[23] Even Disney Is Worried About the High Cost of a Disney Vacation. https://www.msn.com/en-us/travel/news/even-disney-is-worried-about-the-high-cost-of-a-disney-vacation/ar-AA1yFRdv

[24] Naor, P.: The regulation of queue size by levying tolls. Econometrica: journal of the Econometric Society, 15–24 (1969)

[25] Debo, L.G., Toktay, L.B., Van Wassenhove, L.N.: Queuing for expert services. Management Science **54**(8), 1497–1512 (2008)

[26] Borgs, C., Chayes, J.T., Doroudi, S., Harchol-Balter, M., Xu, K.: The optimal admission threshold in observable queues with state dependent pricing. Probability in the Engineering and Informational Sciences **28**(1), 101–119 (2014)

[27] Frazelle, A.E., Katok, E.: Paid priority in service systems: Theory and experiments. Manufacturing & Service Operations Management **26**(2), 775–795 (2024)

[28] Çil, E.B., Karaesmen, F., Örmeci, E.L.: Dynamic pricing and scheduling in a multi-class single-server queueing system. Queueing Systems **67**(4), 305–331 (2011)

[29] Maglaras, C.: Revenue management for a multiclass single-server queue via a fluid model analysis. Operations Research **54**(5), 914–932 (2006)

[30] Stanford, D.A., Taylor, P., Ziedins, I.: Waiting time distributions in the accumulating priority queue. Queueing Systems **77**, 297–330 (2014)

[31] Jiang, Y., Tham, C.-K., Ko, C.-C.: A probabilistic priority scheduling discipline for multi-service networks. Computer Communications **25**(13), 1243–1254 (2002)

[32] Tham, C.-K., Yao, Q., Jiang, Y.: Achieving differentiated services through multi-class probabilistic priority scheduling. Computer Networks **40**(4), 577–593 (2002)

[33] Furth, P.G., Muller, T.H.: Conditional bus priority at signalized intersections: better service with less traffic disruption. Transportation research record **1731**(1), 23–30 (2000)

[34] Garrow, M., Machemehl, R.: Development and evaluation of transit signal priority strategies. Journal of Public Transportation **2**(2), 65–90 (1999)

[35] Ding, Q., Kouvelis, P., Milner, J.: Inventory rationing for multiple class demand under continuous review. Production and Operations Management **25**(8), 1344–1362 (2016)

[36] Moon, I., Kang, S.: Rationing policies for some inventory systems. Journal of the Operational Research Society **49**(5), 509–518 (1998)

[37] Ayesta, U.: A unifying conservation law for single-server queues. Journal of Applied Probability **44**(4), 1078–1087 (2007). Accessed 2024-07-26

[38] Pareto, V.: Cours D'économie Politique vol. 1. Librairie Droz, Geneva, Switzerland (1964)

[39] Jones, C.I.: Pareto and Piketty: The macroeconomics of top income and wealth inequality. Journal of Economic Perspectives **29**(1), 29–46 (2015)

[40] Our Guide to Disney Lightning Lane Prices. https://www.wdw-magazine.com/lightning-lane-multi-pass-cost/

[41] Harchol-Balter, M.: Performance Modeling and Design of Computer Systems: Queueing Theory in Action. Cambridge University Press, Cambridge (2013)

A The number of jobs in a busy period

We present a lemma characterizing the number of jobs in a busy period. The proof is similar to the well-known formula of the length of a busy period ([41, Chapter 27]). We use $\widetilde{X}(s) = \mathbf{E}\left[e^{-sX}\right]$ to denote the Laplace transform of random variable X and $\widehat{Y}(z) = \mathbf{E}\left[z^{Y}\right]$ to denote the z-transform of random variable Y.

Lemma 4. *Let N denote the number of jobs in a busy period. Then*

$$\widetilde{N}(s) = e^{-s} \cdot \widetilde{S}(\lambda - \lambda\widetilde{N}(s)).$$

Proof. Let $N(x)$ denote the number of jobs in a busy period started by a job with size x. Then we have the recursive formula

$$N(x) = 1 + \sum_{i=1}^{A_x} N^{(i)},$$

where A_x is the number of jobs arriving during a period of time of length x.

Thus we have that

$$\begin{aligned}\widetilde{N(x)}(s) &= e^{-s} \cdot \widehat{A_x}\left(\widetilde{N}(s)\right) \\ &= e^{-s} \cdot e^{-\lambda x(1-\widetilde{N}(s))}.\end{aligned}$$

Integrating over x gives the proof:

$$\begin{aligned}\widetilde{N}(s) &= \int_0^\infty \widetilde{N(x)}(s) \cdot f_S(x)dx \\ &= \int_0^\infty e^{-s} \cdot e^{-\lambda x(1-\widetilde{N}(s))} \cdot f_S(x)dx \\ &= e^{-s} \cdot \widetilde{S}(\lambda - \lambda\widetilde{N}(s))\end{aligned}$$

□

B Achievability Region of Hybrid

In this section we mathematically quantify the *achievability region* of Hybrid, i.e., the region of all permissible waiting time pairs for queue 1 and queue 2, namely $(\mathbf{E}[W_1], \mathbf{E}[W_2])$. The goal is to prove the region in Figure 1.

We start in Section B.1 by introducing notation and definitions. Then in Section B.2 and Section B.3, we derive regions in Figure 1.

B.1 Definitions and Notations

The definition of the Hybrid policy is in Section 3.1.

Let Prio(1;2) (respectively, Prio(2;1)) denote the strict priority policy where queue 1 (respectively, queue 2) customers have priority. Note that Prio(1;2) and Prio(2;1) are special cases of Hybrid priority, where $q = 1$ or $q = 0$.

For the purpose of this section, our model is just two queues, the first with arrival rate λ_1 and the second with arrival rate λ_2, where both arrival processes are Poisson. There is a single server which serves customers from both queues in a non-preemptive fashion, according to Hybrid(q). We assume that the service requirement of customers in queue 1 is drawn from distribution S_1 and that of customers in queue 2 is drawn from S_2. Let $\mu_1 = \frac{1}{\mathbf{E}[S_1]}, \mu_2 = \frac{1}{\mathbf{E}[S_2]}$ and $\rho_1 = \frac{\lambda_1}{\mu_1}$, $\rho_2 = \frac{\lambda_2}{\mu_2}$. Define $\lambda := \lambda_1 + \lambda_2, \rho = \rho_1 + \rho_2$. Define S to be the service requirement of a customer, i.e.,

$$S = \begin{cases} S_1 & \text{with probability } \frac{\lambda_1}{\lambda} \\ S_2 & \text{otherwise.} \end{cases}$$

We make a mild assumption that $\mathbf{E}\left[S^3\right]$ is finite, which is a technicality which will be needed in Lemma 6. We define $\mathbf{E}[S_e] := \frac{\mathbf{E}[S^2]}{2\mathbf{E}[S]}$ to be the expected excess.

B.2 Strict Priority

B.2.1 Expected Waiting Times under Strict Priority:

The mean waiting times for queues 1 and 2, namely $\mathbf{E}[W_1]$ and $\mathbf{E}[W_2]$, are well-known, under strict non-preemptive priority, see e.g. [41, p. 502]:

$$\mathbf{E}[W_1]^{Prio(1;2)} = \frac{\rho \mathbf{E}[S_e]}{1-\rho_1} \tag{17}$$

$$\mathbf{E}[W_2]^{Prio(1;2)} = \frac{\rho \mathbf{E}[S_e]}{(1-\rho_1)(1-\rho)}. \tag{18}$$

B.2.2 Achievability Region of Strict Priority:

Proposition 5 derives the achievability region for strict priority, and the blue shaded area in Figure 1 depicts the stability region for Prio(1;2) and Prio(2;1) graphically. As

you can see, the achievability regions for Prio(1;2) and Prio(2;1) have narrow tornado-like shapes. The proof of Proposition 5 follows from (17) and (18). While the proof is straightforward, the picture of the narrow tornado-shape regions demonstrates the limitation of strict priorities, and is not prominent in the literature.

Proposition 5 (Achievability Region of Strict Priority)**.** *A point* $(x = \mathbf{E}[W_1], y = \mathbf{E}[W_2])$ *in the waiting time plane lies in the achievability region of strict priority iff*

$$x \geqslant \frac{y\mathbf{E}[S_e]}{y + \mathbf{E}[S_e]} \qquad \textit{and} \qquad y \geqslant \frac{x^2}{\mathbf{E}[S_e]} + x.$$

Proof. Suppose point (x, y) is achievable. then

$$x = \mathbf{E}[W_1]^{Prio(1;2)} = \frac{\rho\mathbf{E}[S_e]}{1-\rho_1} \tag{19}$$

$$y = \mathbf{E}[W_2]^{Prio(1;2)} = \frac{\rho\mathbf{E}[S_e]}{(1-\rho_1)(1-\rho)}. \tag{20}$$

Taking (19) and dividing it by (20) yields $(1-\rho) = \frac{x}{y}$ and thus

$$\rho = 1 - \frac{x}{y}. \tag{21}$$

Returning to (19), we have that $1 - \rho_1 = \frac{\rho\mathbf{E}[S_e]}{x}$ and thus

$$\rho_1 = 1 - \frac{\rho\mathbf{E}[S_e]}{x}. \tag{22}$$

If we now substitute in (21) into (22), we get:

$$\rho_1 = 1 - \frac{\mathbf{E}[S_e]}{x} + \frac{\mathbf{E}[S_e]}{y}. \tag{23}$$

We now use the fact that $\rho_1 \geqslant 0$ and the fact that $\rho_2 = \rho - \rho_1 \geqslant 0$ to complete the proof. Specifically, setting $\rho_1 \geqslant 0$, from (23) we can equivalently write:

$$1 - \frac{\mathbf{E}[S_e]}{x} + \frac{\mathbf{E}[S_e]}{y} \geqslant 0,$$

which solves to

$$x \geqslant \frac{y\mathbf{E}[S_e]}{y + \mathbf{E}[S_e]}.$$

Moreover, from (21) and (23) we know that $0 \leqslant \rho_2 = \rho - \rho_1$ is equivalent to:

$$0 \leqslant 1 - \frac{x}{y} - \left(1 - \frac{\mathbf{E}[S_e]}{x} + \frac{\mathbf{E}[S_e]}{y}\right) = \frac{\mathbf{E}[S_e]}{x} - \frac{x + \mathbf{E}[S_e]}{y}.$$

Solving this gives us

$$y \geqslant x^2/\mathbf{E}[S_e] + x.$$

Thus $\rho_1, \rho_2 \geqslant 0$ is equivalent to the two inequalities in the theorem. This completes the proof. □

B.3 Hybrid Priority

B.3.1 Expected Waiting Times under Hybrid

It is not known how to derive the waiting time under Hybrid(q) for a particular value of q. What makes analyzing Hybrid(q) difficult is that the state space for Hybrid(q) is infinite in 2 dimensions (one needs to track both the number of jobs in queue 1 and in queue 2). While all priority systems have a 2D-infinite state space, in the case of Prio(1;2) or Prio(2;1), we can use a "tagged job method" to derive the mean waiting time for each queue, [41]. Unfortunately, Hybrid(q) does not lend itself to such a tagged job analysis.

Fortunately, to derive the achievability region of Hybrid it suffices to understand the range of waiting times spanned by Hybrid(q), and this we derive in Proposition 6

below. The theorem can be summarized by Figure 7, which shows that Hybrid(q) spans the full range from Prio(1;2) to Prio(2;1) as q runs from 0 to 1. Note that similar theorem holds for the partial priority policy in [7], and Hybrid here serves as a more practical partial priority which enjoys the same theoretical property.

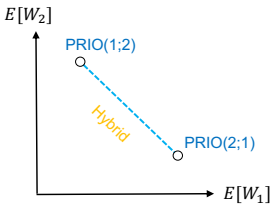


Fig. 7: Hybrid spans the whole segment as q ranges from 0 to 1 for a given (λ_1, λ_2).

Proposition 6 (Hybrid range)**.** *For any pair* (λ_1, λ_2) *such that* $\lambda_1 > 0, \lambda_2 > 0, \rho < 1$, *and any* $0 \leqslant q \leqslant 1$, *there exists an* $\alpha \in [0, 1]$ *s.t.:*

$$\begin{bmatrix} \mathbf{E}[W_1]^{Hybrid(q)} \\ \\ \mathbf{E}[W_2]^{Hybrid(q)} \end{bmatrix} = \alpha \begin{bmatrix} \mathbf{E}[W_1]^{Prio(1;2)} \\ \\ \mathbf{E}[W_2]^{Prio(1;2)} \end{bmatrix} + (1-\alpha) \begin{bmatrix} \mathbf{E}[W_1]^{Prio(2;1)} \\ \\ \mathbf{E}[W_2]^{Prio(2;1)} \end{bmatrix}, \quad (24)$$

and vice versa, i.e., for any $\alpha \in [0, 1]$ *there exists a* $q \in [0, 1]$ *such that (24) is satisfied.*

Proof. The proof consists of two parts.

The first part is to show that the waiting time pair of Hybrid(q) is a linear combination of $\left(\mathbf{E}[W_1]^{\text{Prio}(1;2)}, \mathbf{E}[W_2]^{\text{Prio}(1;2)}\right)$ and $\left(\mathbf{E}[W_1]^{\text{Prio}(2;1)}, \mathbf{E}[W_2]^{\text{Prio}(2;1)}\right)$. This is proved by leveraging the well-known conservation law ([37]):

$$\text{constant} = \rho_1 \mathbf{E}[W_1]^{Hybrid(q)} + \rho_2 \mathbf{E}[W_2]^{Hybrid(q)}. \quad (25)$$

We give a short explanation of why equation (25) holds. Let N_1 and N_2 denote the number of queue 1 and queue 2 customers in the system, respectively. Then

$$\begin{aligned}
\mathbf{E}\left[\text{total work in system}\right] &= \mathbf{E}\left[\text{work in queue}\right] + \mathbf{E}\left[\text{remaining work in service}\right] \\
&= \mathbf{E}\left[N_1\right]\mathbf{E}\left[S_1\right] + \mathbf{E}\left[N_2\right]\mathbf{E}\left[S_2\right] + \rho\mathbf{E}\left[S_e\right] \\
&= \lambda_1\mathbf{E}\left[W_1\right]\mathbf{E}\left[S_1\right] + \lambda_2\mathbf{E}\left[W_2\right]\mathbf{E}\left[S_2\right] + \rho\mathbf{E}\left[S_e\right] \\
&= \rho_1\mathbf{E}\left[W_1\right] + \rho_2\mathbf{E}\left[W_2\right] + \rho\mathbf{E}\left[S_e\right].
\end{aligned}$$

Since Hybrid(q) is work-conserving, we know that $\mathbf{E}\left[\text{total work in system}\right]$ is a constant with respect to q. This gives a proof for (25). From (25) it is immediate to see that the waiting time pair under Hybrid priority lies on a straight line. Notice that $\left(\mathbf{E}\left[W_1\right]^{\text{Prio}(1;2)}, \mathbf{E}\left[W_2\right]^{\text{Prio}(1;2)}\right)$ and $\left(\mathbf{E}\left[W_1\right]^{\text{Prio}(2;1)}, \mathbf{E}\left[W_2\right]^{\text{Prio}(2;1)}\right)$ are extreme cases of the waiting time pair under Hybrid priority, we know that any waiting time pair under Hybrid priority must lie on this line segment.

The second part of the proof shows that $\mathbf{E}\left[W_1\right]^{\text{Hybrid}(q)}$ is continuous with respect to q. We prove this by a sample path argument.

A sample path ω when implementing Hybrid(q) policy consists of two parts: (i) the arrival sequence of customers $\mathcal{A} = (j_1, j_2, ...)$ where each customer is specified with the arrival time and service time; (ii) a sequence of Unif$(0,1)$ variables $\mathcal{U} = u_1, u_2, ...$: the i^{th} coin flip when implementing Hybrid(q) policy is heads if $u_i < q$ and is tails otherwise. Now we compare the average waiting time of queue 1 customers under Hybrid(q) and Hybrid($q+\epsilon$) under the same sample path.

Note that Hybrid(q) is work conserving, thus the busy periods division is the same as that under FCFS policy, i.e., two customers are in the same busy period under Hybrid(q) if and only if they are in the same busy period under FCFS policy. Thus the busy periods division is also the same under Hybrid(q) and Hybrid($q+\epsilon$) policies,

and the length of a busy period follows the well known formula ([41, Chapter 27]):

$$\widetilde{B}(s) = \widetilde{S}(s + \lambda - \lambda\widetilde{B}(s)), \tag{26}$$

where $\widetilde{X}(s)$ is the Laplace transform of random variable X. Similarly, let N denote the number of jobs in a busy period. By Lemma 4, its transform satisfies the following equation:

$$\widetilde{N}(s) = e^{-s} \cdot \widetilde{S}(\lambda - \lambda\widetilde{N}(s)). \tag{27}$$

Define a coin flip to be *affected* if it is heads under Hybrid$(q + \epsilon)$ but tails under Hybrid(q). Note that the event that a coin flip is different under Hybrid(q) and Hybrid$(q + \epsilon)$ is equivalent to the event that the corresponding Unif$(0, 1)$ random variable lies in $(q, q + \epsilon]$. Thus we know that a coin flip is affected with probability ϵ.

Now we look at a busy period $\mathcal{B}^*$. Let B^* denote its length and N^* denote the number of jobs in $\mathcal{B}^*$. Recall that when implementing Hybrid(q), a coin is flipped whenever the server is free and there is a job waiting. Thus running jobs in $\mathcal{B}^*$ uses N^* coin flips (corresponding to N^* random variables in $\mathcal{U}$). Let N_a^* denote the number of affected coin flips among the N^* coin flips.

Now for any customer j^* in the busy period $\mathcal{B}^*$, let $W^q(*)$ denote its waiting time under Hybrid(q) policy. Note that if $N_a^* = 0$, all customers in $\mathcal{B}^*$ are served in the same order under Hybrid(q) and Hybrid$(q + \epsilon)$, which means $W^q(*) = W^{q+\epsilon}(*)$. On the other hand, the customer j^* is delayed for at most B^* time since B^* is the length of the whole busy period, Thus $W^q(*) \leqslant B^*$, which indicates that

$$W^q(*) - W^{q+\epsilon}(*) \leqslant B^*.$$

Moreover, there are N^* coin flips in $\mathcal{B}^*$ and each of them has an independent ϵ probability to be affected. Thus we know that

$$\mathbb{P}\left[N_a^* = 0 \mid \mathcal{B}^*\right] = (1-\epsilon)^{N^*}.$$

Thus we have that

$$\begin{aligned}
\mathbf{E}\left[W^q(*) - W^{q+\epsilon}(*) \mid \mathcal{B}^*\right] &\leqslant \mathbb{P}\left[N_a^* > 0 \mid \mathcal{B}^*\right] \cdot B^* \\
&= \left(1 - (1-\epsilon)^{N^*}\right) \cdot B^* \\
&\leqslant (1 - (1 - N^* \cdot \epsilon)) \cdot B^* \\
&= \epsilon \cdot N^* \cdot B^*.
\end{aligned}$$

Finally, we have the inequality

$$\begin{aligned}
&\mathbf{E}\left[W_1\right]^{\text{Hybrid}(q)} - \mathbf{E}\left[W_1\right]^{\text{Hybrid}(q+\epsilon)} \\
&= \mathbf{E}\left[W^q(*) - W^{q+\epsilon}(*)\right] \\
&= \mathbf{E}_{\mathcal{B}^*}\left[\mathbf{E}\left[W^q(*) - W^{q+\epsilon}(*) \mid \mathcal{B}^*\right] \cdot \frac{B^*}{\mathbf{E}\left[B\right]}\right] && \text{Inspection Paradox} \\
&\leqslant \mathbf{E}_{\mathcal{B}^*}\left[\epsilon \cdot N^* \cdot B^* \cdot \frac{B^*}{\mathbf{E}\left[B\right]}\right] && \text{Inequality above} \\
&= \frac{\epsilon}{\mathbf{E}\left[B\right]} \mathbf{E}\left[N \cdot B^2\right].
\end{aligned}$$

Moreover, we know that

$$\mathbf{E}\left[N \cdot B^2\right] \leqslant \frac{1}{3}\mathbf{E}\left[N^3 + 2B^3\right] = \frac{1}{3}\left(\mathbf{E}\left[N^3\right] + 2\mathbf{E}\left[B^3\right]\right).$$

By equations (26), (27) and the assumption that $\mathbf{E}\left[S^3\right]$ exists, we know that $\mathbf{E}\left[N^3\right]$ and $\mathbf{E}\left[B^3\right]$ are finite. Thus we have the proof that $\mathbf{E}\left[W_1\right]^{\text{Hybrid}(q)}$ is continuous with q. □

B.3.2 Hybrid's achievability region is vast

We now want to argue that Hybrid's achievability region includes the entire region between (and including) the achievability regions of Prio(1;2) and Prio(2;1). We will use $\mathcal{A}_{hybrid}$ to denote the union of the yellow and blue regions in Figure 1. In Proposition 7, we will show that Hybrid covers every point in $\mathcal{A}_{hybrid}$.

Proposition 7 (Achievability Region for Hybrid Queue)**.** *The Achievability Region for Hybrid is* $\mathcal{A}_{hybrid}$.

Proof. Note that since the waiting time for Hybrid lies on the line segment determined by the Strict Priority policies, it is obvious that Hybrid cannot achieve any point outside of $\mathcal{A}_{hybrid}$.

We now show that the achievability region for Hybrid is all of $\mathcal{A}_{hybrid}$. We prove this by contradiction. Suppose there is a point $G = (x_0, y_0) \in \mathcal{A}_{hybrid}$, shown in Figure 8, which is not achievable by Hybrid. Define a set

$$S := \{(x, y) \mid x \leqslant x_0 \ \& \ y \leqslant y_0 \ \& \ (x, y) \text{ is achievable for Hybrid}\}.$$

Since $(0, 0) \in S$, we know that S is not the empty set. Define $P = (x_1, y_1)$ to be the point in S which is closest to point G (we break ties arbitrarily).

By definition, there exist parameters $(\lambda_1, \lambda_2, q)$ for Hybrid that satisfy the waiting times specified by P. Now, consider the waiting times for Prio(1;2) and Prio(2;1) under those same arrival rates. Let

$$A = \left(\mathbf{E}\left[W_1 \;\middle|\; \lambda_1, \lambda_2\right]^{Prio(1;2)}, \mathbf{E}\left[W_2 \;\middle|\; \lambda_1, \lambda_2\right]^{Prio(1;2)}\right),$$

$$B = \left(\mathbf{E}\left[W_1 \;\middle|\; \lambda_1, \lambda_2 \right]^{Prio(2;1)}, \mathbf{E}\left[W_2 \;\middle|\; \lambda_1, \lambda_2 \right]^{Prio(2;1)} \right).$$

Then, by Proposition 6, we know that P is on the line segment $\overline{AB}$.

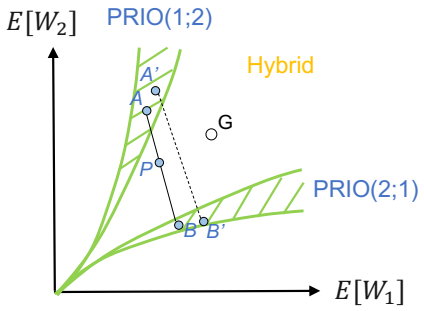


Fig. 8: Illustration for G, P, A, B for the proof of .

Case 1: $x_1 < x_0$ & $y_1 < y_0$. Note that $\lambda_1 + \lambda_2 < 1$. Therefore we can define

$$\lambda_1' = \lambda_1 + \epsilon \quad \text{and} \quad \lambda_2' = \lambda_2,$$

which correspond to new points in Figure 8: A' and B'. From the monotonicity of the waiting times in Prio(1;2) and Prio(2;1), we know that both coordinates of A' exceed those of A; likewise, both coordinates of B' exceed those of B. Hence, the line segment $\overline{A'B'}$ is slightly closer to G. Thus we can pick a node P' on $\overline{A'B'}$ which is closer to G than P. By Proposition 6, $P' \in S$, which is contradictory with how we selected P.

Case 2: $x_1 = x_0$ or $y_1 = y_0$. WLOG suppose $x_1 = x_0$. Since

$$\mathbf{E}\left[W_2 \;\middle|\; \lambda_1, \lambda_2 \right]^{Prio(1;2)} > \mathbf{E}\left[W_2 \;\middle|\; \lambda_1, \lambda_2 \right]^{Prio(2;1)},$$

we know that the line $\overline{AB}$ has a negative slope. Let the node P' be the node on the line segment $\overline{AB}$ whose y-coordinate is $y_1 + \epsilon$. By Proposition 6, $P' \in S$. But P' is closer to G than P, which is contradictory with how we selected P. □

C Proof for Variant of Model from Section 4.3

The optimization of this model is:

$$\boxed{\begin{array}{ll} \underset{\lambda_1, q, \$}{\text{maximize}} & \$ \cdot \lambda_1 \\ \text{s.t.} & \dfrac{\lambda_1}{\lambda} = \bar{F}_C\left(\dfrac{\$}{\mathbf{E}[W_2] - \mathbf{E}[W_1]}\right), \\ & \mathbf{E}[W_2] \leqslant \overline{W}, \\ & \$ \leqslant \bar{\$}, \\ & 0 \leqslant q \leqslant 1. \end{array}} \tag{28}$$

By exactly the same computation in Lemma 1, this optimization problem is equivalent to

$$\boxed{\begin{array}{ll} \underset{\lambda_1, q}{\text{maximize}} & \$ \cdot \lambda_1 \\ \text{s.t.} & \$ = \theta(\lambda_1, q), \\ & \mathbf{E}[W_2] \leqslant \overline{W}, \\ & \$ \leqslant \bar{\$}, \\ & 0 \leqslant q \leqslant 1. \end{array}} \tag{29}$$

In contrast to the fact that there may not be a feasible solution under strict priority, a solution always exists under Hybrid.

Lemma 5. *There always exists a solution under Hybrid (or partial priority).*

Proof. The proof is straightforward. For any λ_1, set q such that $\mathbf{E}[W_1] = \mathbf{E}[W_2] = w_{FCFS}$. Then $\theta(\lambda_1, q) = 0$ and all restrictions are satisfied. □

An analogue of Lemma 2 is given below:

Lemma 6. *At least one of the following conditions must hold:*

1. $\mathbf{E}[W_2^*] = \overline{W}, \theta^* = \bar{\$}$.

2. $q^* = 1$.

Proof. We prove by contradiction. Suppose both of them are false. Then we have $q^* < 1$. We discuss different cases and show that all of them are contradictory.

We perturb λ_1^* and q^* in all cases to prove contradictions. This approach is feasible because $\lambda_1^* \in (0, \lambda)$ and $q^* \in (0, 1)$. To justify this, note that if $\lambda_1^* = 0$, the revenue would be zero which is clearly suboptimal. If $\lambda_1^* = \lambda$, then $\theta^* = 0$ because $g(1) = 0$, which also leads to zero revenue. If $q^* = 0$, then $\theta(\lambda_1^*, q^* = 0) < 0$ because $\mathbf{E}[W_2 \mid \lambda_1^*, q^* = 0] < w_{FCFS}$. Finally, if $q^* = 1$, condition (2) above holds.

Case 1: $\mathbf{E}[W_2^*] < \overline{W}$. In this case, we increase λ_1^* to $\lambda_1' = \lambda_1^* + \epsilon$. Now we perturb q^* to q' such that

$$\theta(\lambda_1', q') = \theta^*. \tag{30}$$

This can be done when ϵ is small enough by the continuity of θ. This change of value can still guarantee that $\mathbf{E}[W_2'] \leqslant \overline{W}$ if ϵ is small enough. In this way, λ_1 is increased and the price is the same, which increases the revenue.

Case 2: $\mathbf{E}[W_2^*] = \overline{W}, \theta^* < \bar{\$}$. In this case, the revenue is

$$\text{Revenue}^* = \theta^* \cdot \lambda_1^* = \lambda \cdot g\left(\frac{\lambda_1^*}{\lambda}\right) \cdot \left(\mathbf{E}[W_2^*] - w_{FCFS}\right).$$

Thus we can decrease λ_1^* to $\lambda_1' = \lambda_1^* - \epsilon$ and increase q^* to q' such that $\mathbf{E}[W_2 \mid \lambda_1', q'] = \mathbf{E}[W_2^*]$. Make ϵ small enough such that $\theta(\lambda_1', q') \leqslant \bar{\$}$ still holds. In this way, the revenue is also increased because g is decreasing with respect to λ_1.

Combining those two cases gives the proof. □

Lemma 7 (When does Hybrid help?)**.** *Hybrid (or partial priority) helps increase the revenue compared with strict priority if and only if*

$$\theta(\widehat{\lambda_1}, q = 1) > \bar{\$}. \tag{31}$$

Proof. We first prove that $\theta(\widehat{\lambda_1}, q = 1) > \overline{\$}$ is a *necessary condition* for Hybrid to help.

Suppose by contradiction that $\theta(\widehat{\lambda_1}, q = 1) \leqslant \overline{\$}$ and assume that Hybrid still helps. This is saying that we obtain the optimal revenue with $q^* < 1$. By Lemma 6, we have that $\mathbf{E}[W_2^*] = \overline{W}$ and $\theta^* = \overline{\$}$. Then

$$\text{Revenue}^* = \theta^* \cdot \lambda_1^* = \lambda \cdot g\left(\frac{\lambda_1^*}{\lambda}\right) \cdot (\mathbf{E}[W_2^*] - w_{FCFS}) = \lambda \cdot g\left(\frac{\lambda_1^*}{\lambda}\right) \cdot \left(\overline{W} - w_{FCFS}\right).$$

On the other hand, if we set $\lambda_1 = \widehat{\lambda_1}$ under strict priority, the revenue obtained is

$$\text{Revenue}' = \theta(\widehat{\lambda_1}, q = 1) \cdot \widehat{\lambda_1} = \lambda \cdot g\left(\frac{\widehat{\lambda_1}}{\lambda}\right) \cdot \left(\overline{W} - w_{FCFS}\right).$$

Since the optimal $q^* < 1$, we know that

$$\text{Revenue}^* > \text{Revenue}',$$

which indicates $\lambda_1^* < \widehat{\lambda_1}$ because g is decreasing.

But this leads to the contradiction:

$$\overline{W} = \mathbf{E}[W_2 \mid \lambda_1^*, q^*] < \mathbf{E}[W_2 \mid \lambda_1^*, q = 1] \leqslant \mathbf{E}\left[W_2 \mid \widehat{\lambda_1}, q = 1\right] = \overline{W}.$$

We now prove that $\theta(\widehat{\lambda_1}, q = 1) > \overline{\$}$ is a sufficient condition for Hybrid to help. In this case, the optimal revenue under strict priority is smaller than $\widehat{\lambda_1} \cdot \overline{\$}$ because under strict priority, ensuring $\mathbf{E}[W_2] \leqslant \overline{W}$ requires $\lambda_1 \leqslant \widehat{\lambda_1}$, and $\$ \leqslant \overline{\$}$ because of the price cap.

Now we can pick a small $\epsilon > 0$ and make $q' = 1 - \epsilon$. Accordingly we can increase $\widehat{\lambda_1}$ to λ_1' such that $\mathbf{E}[W_2 \mid \lambda_1', q'] = \mathbf{E}\left[W_2 \mid \widehat{\lambda_1}, q = 1\right]$. Pick ϵ small enough such that $\theta(\lambda_1', q') \geqslant \overline{\$}$ still holds. Now keep decreasing q' until $\theta(\lambda_1', q') = \overline{\$}$. In this way,

the revenue is $\bar{\$} \cdot \lambda_1' > \bar{\$} \cdot \widehat{\lambda_1}$ which is larger than the optimal revenue under strict priority. □

Theorem 3 follows by simplifying the condition in Lemma 7. The simplification is exactly the same as that in the proof of Thereom 1. We now present the proof of Theorem 4.

Proof of Theorem 4. By Theorem 3, condition (15) is equivalent to condition (31) in Lemma 7, which is $\theta(\widehat{\lambda_1}, q = 1) > \bar{\$}$.

To prove the first argument, since $\theta(\widehat{\lambda_1}, q = 1) > \bar{\$}$ and there is no $\lambda_1 \leqslant \widehat{\lambda_1}$ such that $\theta(\lambda_1, q = 1) = \bar{\$}$, we have that for any $\lambda_1 \leqslant \widehat{\lambda_1}$, $\theta(\lambda_1, q = 1) > \bar{\$}$. This means there is no solution under strict priority. Since there always exists a solution under Hybrid (Lemma 5), we know that Hybrid beats strict priority.

To prove the second argument, we first notice that under strict priority, the optimal revenue is $\lambda_1^{q=1} \cdot \bar{\$}$, which is achieved when $\lambda_1 = \lambda_1^{q=1}$. This is because by the continuity of θ, any λ_1 such that $\theta(\lambda_1, 1) \leqslant \bar{\$}$ must satisfy $\lambda_1 \leqslant \lambda_1^{q=1}$, and the price is capped by $\bar{\$}$. This gives that $\text{Revenue}(q = 1) = \lambda_1^{q=1} \cdot \bar{\$}$.

On the other hand, by Theorem 3, we know that the optimal $q^* < 1$. Thus by Lemma 6 we know that

$$\theta^* = \bar{\$}, \quad \mathbf{E}\left[W_2^*\right] = \overline{W}. \tag{32}$$

This shows that the improvement ratio is

$$\text{Improvement Ratio} = \frac{\bar{\$} \cdot \lambda_1^*}{\bar{\$} \cdot \lambda_1^{q=1}} = \frac{\lambda_1^*}{\lambda_1^{q=1}}.$$

The expression for λ_1^* can be solved from (32), which is equivalent to

$$\frac{\lambda}{\lambda_1^*} \cdot g\left(\frac{\lambda_1^*}{\lambda}\right) = \frac{\bar{\$}}{\overline{W} - w_{FCFS}}.$$

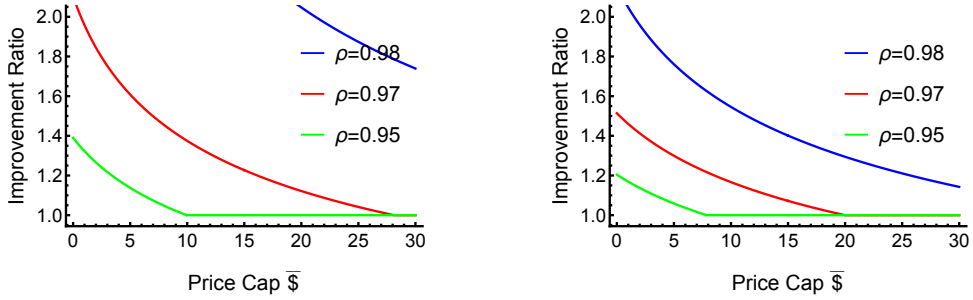
Note that the left hand side of the equation is continuous, monotone with λ_1^*, and ranges from ∞ to 0 when λ_1^* takes from 0 to λ, thus there exists a unique solution of λ_1^*.

Finally, the formula for the improvement amount is given by

$$\text{Improvement Amount} = (\text{Improvement Ratio} - 1) \cdot \text{Revenue}(q = 1).$$

□

D Additional Figures for Section 5



(a) Max Wait $\overline{W} = 30$.

(b) Max Wait $\overline{W} = 45$.

Fig. 9: The improvement ratio does not change a lot compared with Figure 4 when α drops. In this set of experiments, we set $\alpha = 1.1$, but keep the other parameters the same as Figure 4, namely $S = 1$ (deterministic), and $\overline{W}$ takes on the values shown in (a) and (b).